Multi-lingual and Cross-genre Discourse Unit Segmentation

Peter Bourgonje and Robin Schäfer

Applied Computational Linguistics
University of Potsdam / Germany
firstname.lastname@uni-potsdam.de

Abstract

We describe a series of experiments applied to data sets from different languages and genres annotated for coherence relations according to different theoretical frameworks. Specifically, we investigate the feasibility of a unified (theory-neutral) approach toward discourse segmentation; a process which divides a text into minimal discourse units that are involved in some coherence relation. We apply a RandomForest and an LSTM based approach for all data sets, and we improve over a simple baseline assuming simple sentence or clause-like segmentation. Performance however varies a lot depending on language, and more importantly genre, with f-scores ranging from 73.00 to 94.47.

1 Introduction

The last few decades have seen several different theories and frameworks being proposed for the task of *discourse processing*, or *discourse parsing*; the analysis and (automatic) extraction of coherence relations from a text. Among the most popular approaches are Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), the Penn Discourse Treebank (PDTB) (Prasad et al., 2008), Segmented Discourse Representation Theory (SDRT) (Asher et al., 2003) and the Cognitive approach to Coherence Relations (CCR) (Sanders et al., 1992). While each of these approaches may serve a different purpose or have a specific focus, to a certain extent they all rely on segmenting texts into segments that express specific propositions which make up the arguments or components of some relation. The 2019 DISRPT workshop aims to contribute to a shared understanding of coherence relations by providing training and evaluation data from several available treebanks in the RST, SDRT and PDTB formalisms. Because each of these formalisms have their specific characteristics for the various stages of analyses (i.e. differences in segmentation, relation inventory, flat or tree-like representations, etc.) the shared task[1] accompanying the workshop is meant to promote the design of flexible methods for dealing with these differences. The focus is on the first (and comparably easiest) step in the process; segmenting a text into minimal units, as a standard for discourse segmentation would, in addition to a better general understanding, allow treebanks or resources annotated according to one theoretical framework to help in (manually or automatically) annotating data according to other frameworks. In this paper we describe a set of experiments using the collection of data sets provided in the context of the shared task, including nine different languages and a variety of genres.

The rest of the paper is structured as follows: Section 2 describes related work in this direction. Section 3 describes the three formalisms that are present among the data sets and the data sets themselves. Section 4 describes our approach toward the segmentation task. Sections 5 and 6 present and discuss the results, respectively, and finally, Section 7 sums up our approach and main findings.

2 Related Work

Since the introduction of RST in Mann and Thompson (1988), several discourse parser for English have been proposed ((Soricut and Marcu, 2003), (Hernault et al., 2010), (Ji and Eisenstein, 2014), (Joty et al., 2015)). Additionally, the release of the PDTB (Prasad et al., 2008) helped further enabling machine-learning approaches toward shallow discourse parsing through its relatively large size (compared to RST and also SDRT cor-

[1] https://github.com/disrpt/sharedtask2019

Proceedings of Discourse Relation Parsing and Treebanking (DISRPT2019), pages 105–114
Minneapolis, MN, June 6, 2019.

pora). More recently, the 2015 and 2016 CoNLL shared tasks, following the PDTB framework, sparked interest for the task of shallow discourse parsing, with Wang and Lan (2015) and Oepen et al. (2016) as winning systems, respectively. The tasks featured both English and Chinese discourse parsing. With the generation of several treebanks in other languages over the last decade(s) (see Table 1 in Section 3 for an overview), training and evaluation data became available for several other languages as well (where before systems had to be rule-based, as the one described in Pardo and Nunes (2008)). On the topic of multi-lingual parsing, Braud et al. (2017) describe a cross-lingual approach to RST parsing, using 6 of the 9 corpora used in our experiments, but use language-specific segmenters for the languages they work with (Basque, Dutch, English, German, Spanish and Brazilian Portuguese). Iruskieta et al. (2016) look at a particular kind of segment and detect central units in both Basque and Brazilian Portuguese, where they define central units (CUs) to be units that "(do) not function as satellite of any other unit or text span.". Earlier work on unifying discourse parsing frameworks is described in Rehbein et al. (2016), Benamara and Taboada (2015), Bunt and Prasad (2016), Chiarcos (2014) and Sanders et al. (2018) from a theoretical perspective, and in Demberg et al. (2017) from a practical perspective, but their main focus is on relation senses. Although this presupposes some sort of mapping of units, language- and data-set individual segmentation can be, and in many cases is used. The 2019 DISRPT shared task will undoubtedly generate many more contributions to the segmentation task specifically.

3 Data

The data that is featured in the shared task stems from three different formalisms and covers nine different languages. An overview of the data sets, their formalism and size is shown in Table 1. Note that the indicated number of tokens are for the training and development sets only[2].

The three different formalisms that the treebanks originated from, each have their own conventions, underlying theory and potential application scenarios. While these bridges may be too large to gap for the entire representation of coherence relations, when it comes to just text segmentation, interesting synergies, and perhaps even unified approaches can be explored. In what follows, we will briefly explain the most important specifics with regard to segmentation of each of the three theories featured in the shared task, to conclude with our expectations in terms of overlap when dealing with the sub-task of segmentation alone.

3.1 RST

Introduced by Mann and Thompson (1988), RST aims to represent a text as a single tree structure, in which every single token is included in some elementary discourse unit (EDU) which serves as either a satellite or a nucleus in some relation. EDUs can be sequences of tokens at text level, or can be complex sub-trees which hierarchically represent a larger body of text. In RST, segmentation is an important first step in analysing a text (and consequently generating a tree); before a hierarchy of EDUs can be considered, the EDUs themselves have to be identified, which puts the segmentation task at the center of any RST analysis.

3.2 PDTB

In contrast to RST, in the PDTB framework, which originated as a discourse annotation layer over the Penn Treebank (Miltsakaki et al., 2004), no commitment to the overall structure of the text is made, an approach typically referred to as shallow discourse parsing. Relations between two (often adjacent, but not necessarily so) pieces of text are classified according to a set of relation senses. This is first done by locating explicit connectives and their two arguments (internal, or *arg2* and external, or *arg1*). Subsequently, adjacent sentences inside the same paragraph that are not yet connected through an explicit relation are classified according to an (implicit) relation sense, or as entrel or no-rel (see Prasad et al. (2008) for more details). Segmentation plays a less central role and is somewhat less formally defined. The two arguments of a relation should refer to *propositions* and typically include a finite verb, but under certain circumstances exceptions are made (for nominalized constructions such as "the uprising of the Bolsheviks" for example).

3.3 SDRT

SDRT (Asher et al., 2003) was proposed as an extension to Discourse Representation Theory

[2]The test sets were added only in the final stage of the shared task.

Corpus name	Language	Annotation style	Tokens
RSTBT (Iruskieta et al., 2013)	Basque	RST	28,658
CDTB (Zhou and Xue, 2015)	Chinese	PDTB	63,239
SCTB (Cao et al., 2018)	Chinese	RST	11,067
NLDT (Redeker et al., 2012)	Dutch	RST	21,355
PDTB (Prasad et al., 2008)	English	PDTB	1,100,990
GUM (Zeldes, 2017)	English	RST	82,691
RSTDT (Carlson et al., 2002)	English	RST	184,158
STAC (Asher et al., 2016)	English	SDRT	41,666
ANNODIS (Afantenos et al., 2012)	French	SDRT	25,050
PCC (Stede and Neumann, 2014)	German	RST	29,883
RRST (Toldova et al., 2017)	Russian	RST	243,896
RSTSTB (da Cunha et al., 2011)	Spanish	RST	50,565
SCTB (Cao et al., 2018)	Spanish	RST	12,699
CSTN (Cardoso et al., 2011)	Brazilian Portuguese	RST	51,041

Table 1: Shared task data sets

(Kamp, 1981). By including propositions as variables to reason over and discourse relations to rule out certain antecedents or promote others, it accounts for relations in a text beyond the sentence level (where dynamic semantic approaches often fail). Because our contribution deals with discourse segmentation only, and the two corpora included in this paper that have SDRT annotations both use RST-style EDUs for initial segmentation, the differences between the two theories are irrelevant for the segmentation task at hand.

3.4 Segmentation & Overlap

Segmentation of text into minimal units is not the first step in processing some piece of text in all of the frameworks described above. In PDTB for example, typically explicit signals in the form of connectives are identified first, upon which their arguments are extracted. Subsequently extracting implicit relations more or less means filling in the blanks between explicit relations. In RST and in the two corpora with SDRT annotations, it plays a much more central role, and segmenting a text into EDUs is the first step in constructing a tree for some text. Annotating coherence relations is a time-consuming and difficult task, as is reflected by low inter-annotator agreement scores compared to other NLP tasks, especially when using the RST framework (because of its requirement to end up with one single tree-like representation covering the entire text). As a result, available annotated corpora are relatively small and sparse. For this reason alone, attempting to unify the first, and relatively simple (compared to what follows) step of segmenting some piece of text into minimal units can be very beneficial. Apart from this practical motivation, investigating segmentation characteristics over multiple different frameworks may lead to a broader understanding of the ways meaningful propositions are realised in the languages covered in this shared task. Most of the data in the shared task (i.e. the RST and SDRT data sets from Table 1) is annotated for segment boundaries and in addition is provided with dependency trees which, for most data sets, follows the Universal Dependencies scheme, meaning that we have sentence segmentation, part-of-speech tags and position and function for every word in the dependency tree. For the PDTB data sets, instead of segments (EDUs), connectives were labeled, meaning that the information in this data set is of a very different type. Also, the dependency trees were provided for these data sets. Furthermore, note that we did not have access to the Chinese Discourse Treebank, so though labels were provided in the shared task, we do not apply our methods to this data set[3]. Since we worked on the data sets as they were provided by the organisers, for more specific information related to the data sets we refer the reader either to the corresponding publications included in Table 1, or to the shared task website referred to in Section 1.

[3]In the final stage of the shared task, a Turkish data set was added to which we also had no access.

4 Method

To compare results against a simple, yet for some languages and data sets already relatively effective baseline, we first implement our baseline system which either assumes a segment boundary (segment start) at the beginning of each sentence, or at the beginning of each sentence and after every comma. To give a realistic impression of the performance of the other algorithms, the score for the baseline system in Table 2 represents whichever version scored best. This was the version basically assuming every sentence to be a segment for the *RSTBT*, *PDTB*, *GUM*, *RRST*, *RSTSTB* and Spanish *SCTB* data sets. The version assuming a segment boundary after every comma as well performed better for the Chinese *SCTB*, *NLDT*, *RSTDT*, *STAC*, *ANNODIS* and *CSTN* data sets.

4.1 RandomForest

To improve over the baseline, we try two different approaches. The RandomForest method (based on Scikit-learn (Pedregosa et al., 2011)) uses a combination of information present in the CoNLL-format files of the shared task (i.e. the dependency tree) and augment this, where available, with constituency syntax features. The base set of features we use for all languages consists of the surface form of the word itself; the surface forms of the next and previous word; the distance of this word to its parent in the dependency tree; the function of the parent word; the functions of the previous and next word; the part-of-speech tags (both coarse and fine-grained) tag for the previous, current and next word and the parent; binary features for whether or not the previous, current and parent word are starting with an uppercase character; absolute position in the sentence; relative position in the sentence (absolute position divided by sentence length); whether or not there is a verb ((lowercased) coarse part-of-speech tag starts with a "v") in between the current word and the next punctuation mark[4]. We are using the Stanford CoreNLP lexicalized PCFG parser (Klein and Manning, 2003) to obtain constituency trees for the languages supported (Chinese, English, French, German, and Spanish). For data sets in these languages, we additionally use as features the category of the parent node; the categories of the left and right siblings in the tree; the path to the root node and the compressed path to the root node, where consecutive identical nodes are deleted (i.e. [N → NP → S → S] becomes [N → NP → S]). These features are inspired by the approach of Pitler and Nenkova (2009) for connective disambiguation.

4.2 LSTM

The LSTM-based method (based on Keras (Chollet et al., 2015)) uses a smaller feature set, including the distance to the parent, (grammatical) function of the parent and the current word, the parent's pos-tag and the current word's pos-tag, a binary feature for whether or not the first character of the word is uppercased and the relative position in the sentence. For the encoding of the word itself, we use two different approaches; either we use pre-trained word embeddings (Grave et al., 2018), or we use the embeddings from the corpus itself. The approach with pre-trained embeddings performed better for the *RSTBT*, *NLDT*, *RSTDT*, *RRST*, *RSTSTB*, Spanish *SCTB* and *CSTN* corpora, whereas the approach using the embeddings from the corpus itself performed better for the Chinese *SCTB*, *GUM*, *STAC*, *ANNODIS*, *PDTB* and *PCC* corpora. In general though, the scores for the two LSTM approaches were often very close together.

The results when training on the training and development section of every corpus and evaluating on the test section (as defined by the shared task setup) are shown in Table 2. The baseline rows include results for the baseline approach and the RandomForest rows include results using the above-mentioned feature set with the RandomForest classifier. The LSTM rows show the results for the best scoring LSTM system (either the one with pre-trained embeddings or the embeddings from the corpus itself, as explained above). Note that due to a much larger variation in scores over individual runs, for the LSTM approach (regardless of which one specifically), scores are macro-averaged over 10 runs[5]. Our code is publicly available at *reference_anonymised*.

5 Results

For all data sets, we beat the baseline, be it with a small margin for some (*STAC* and the Spanish *SCTB* for example). We did not check for

[4] Any character in the set {!"#$%&'()*+,-./:;¡=¿?@[\]^_`— }

[5] Except for the *RRST* and *PDTB* corpora. Due to their relatively large size, hence longer processing time, these scores were averaged over 5 runs.

			precision	recall	f1 score
Basque	*RSTBT*	baseline	**98.13**	52.95	68.78
		RandomForest	92.60	61.21	73.71
		LSTM	87.75	**68.63**	**75.95**
Chinese	*SCTB*	baseline	87.23	73.21	79.61
		RandomForest	**88.89**	76.19	**82.05**
		LSTM	77.40	**77.32**	76.74
Dutch	*NLDT*	baseline	87.79	**87.54**	87.66
		RandomForest	**98.04**	86.96	**92.17**
		LSTM	90.00	85.94	86.92
English	*PDTB*	baseline	0.13	0.24	0.02
		RandomForest	**38.80**	**35.74**	**37.21**
		LSTM	9.14	9.46	9.29
	GUM	baseline	**100**	73.98	85.05
		RandomForest	97.76	83.54	**90.10**
		LSTM	93.20	**84.69**	88.41
	RSTDT	baseline	66.16	56.10	60.72
		RandomForest	93.89	87.13	90.38
		LSTM	**94.58**	**90.27**	**92.35**
	STAC	baseline	93.47	**94.80**	94.13
		RandomForest	**98.19**	91.49	**94.47**
		LSTM	95.42	90.45	92.81
French	*ANNODIS*	baseline	**93.63**	69.12	79.53
		RandomForest	93.50	80.44	**86.48**
		LSTM	89.61	**82.31**	85.32
German	*PCC*	baseline	**100**	72.45	84.02
		RandomForest	95.74	**84.01**	**89.49**
		LSTM	92.29	82.41	86.90
Russian	*RRST*	baseline	76.04	49.00	59.60
		RandomForest	82.98	67.02	74.15
		LSTM	**84.48**	**70.05**	**76.42**
Spanish	*RSTSTB*	baseline	**97.36**	64.69	77.73
		RandomForest	93.51	75.88	**83.78**
		LSTM	86.21	**76.21**	79.97
	SCTB	baseline	**97.00**	57.74	72.39
		RandomForest	94.33	**59.52**	**73.00**
		LSTM	68.92	55.60	61.45
Brazilian Portuguese	*CSTN*	baseline	64.47	73.96	68.90
		RandomForest	92.07	78.87	84.96
		LSTM	**92.33**	**82.26**	**86.43**

Table 2: Results for the different data sets

statistical significance, so claiming overall improvement over the baseline may not be justified. While the principle behind EDUs is taken to be language-neutral, it is interesting to see that the operationalisations vary greatly among languages and data sets/domains. This is demonstrated by the fluctuation in the baseline scores; from 59.60 (f1 score) for Russian (*RRST*), to 94.13 for English (*STAC*). For all but *STAC* and *CSTN*, precision is higher than recall (and in general comparatively high), meaning that the lower scoring languages have more EDUs per sentence, or just longer sentences on average. The latter is indeed what we see for Brazilian Portuguese, Russian and Basque, with average sentence lengths of 26.87, 21.85 and 19.93 words per sentence, re-

spectively (compared to average lengths of 14.07, 14.74 and 5.05 for the much higher scoring German, Dutch and English STAC data sets, respectively). Since sentence length is a more informative property of domain than of language[6], this may suggest that a language-wise division is not the ideal one, and perhaps the domain should instead serve as the main indicator for performance. In line with the numbers above, we see that the Brazilian Portuguese, Russian and Basque data sets include scientific writing in their corpora, while the German and Dutch data sets tend more to (popular) news commentary, encyclopedia texts (targeted at the general public instead of scientists) and fund-raising letters and commercial advertisements. The English STAC data set is a domain of its own (in-game chats), with very short average sentences. If one takes the level of experience of the author and targeted reader as indications of text complexity (and also as properties of domain), this is likely to correlate to segmentation agreement figures. Unfortunately, mapping domain and complexity onto some shared dimensional space (allowing correlations to arise) is not straightforward. In addition, the creators of corpora used in our experiments do no use a single, easily unify-able metric to calculate Inter-Annotator Agreement (IAA) for EDU segmentation. We do note however that, again, for the higher scoring corpora, IAA was relatively high; Carlson et al. (2002) note a Kappa of 0.97 for RSTDT, Asher et al. (2016) note an initial agreement of 90% for automatic segmentation in STAC, and segmentation is manually improved after this automatic procedure, and Redeker et al. (2012) note an agreement of 97% for EDU segmentation in NLDT. On the other end of the spectrum, Iruskieta et al. (2013) report an EDU agreement of 81.35% for RSTBT and Toldova et al. (2017) report Krippendorf's α figures of 0.2792, 0.3173 and 0.4965 where they consider figures around 0.8 to be acceptable for RRST.

6 Discussion

Figure 1 plots performance for the RandomForest and LSTM approaches (and the baseline for comparison) on the Y axis (f1 score) and the corpora ordered by size (increasing from left to right) on the X axis, illustrating that there is no clear correlation between corpus size and performance. The largest two corpora by a considerable margin[7] (*RSTDT* and *RRST*) do not score better than many of the other, smaller corpora. Regarding the

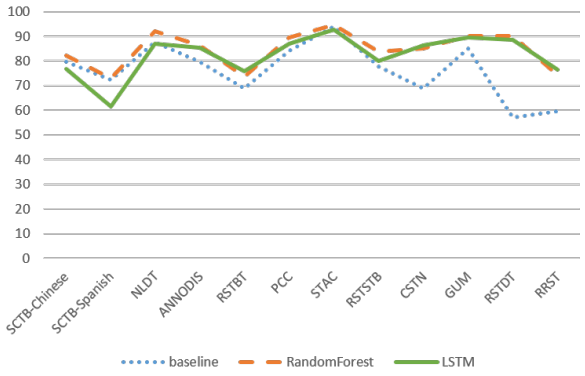

Figure 1: Results for the baseline, RandomForest and LSTM.

RandomForest and LSTM performance, the figure shows that the two come closer together and LSTM outperforms RandomForest on the larger corpora. Overall, RandomForest performs best in 9 data sets, whereas LSTM performs best in 4 cases. The difference however is typically small, and as we did not check for statistical significance, drawing conclusion based on this may not be justified in the cases where the two score close together. The cases where there is a large gap between the baseline and either of the two approaches (CSTN, RSTDT and RRST most notably) all contain (at least a portion of) text from the news domain, but two other corpora containing (a portion of) news text, i.e. PCC and ANNODIS, show much less of a gap. More investigation would be needed for these corpora to find the cause of this gain when using a classifier, compared to the baseline performance.

Figure 2 shows the information gain per feature for the RandomForest classifier for all data sets. Recall that the syntax features based on the constituency tree were not used for all data sets, hence blank for some.

The grammatical function in the dependency tree, (coarse) part-of-speech tag of the parent, position in the sentence, previous word and its part-of-speech tag play an important role for all data sets. For some data sets, the word itself plays

[6]For highly agglutinative languages, depending on tokenisation procedures average sentence lengths may of course be shorter, but given the set of languages here, excluding Chinese, difference in morphology plays a less prominent role.

[7]Excluding the *PDTB*, as explained in the remainder of this section.

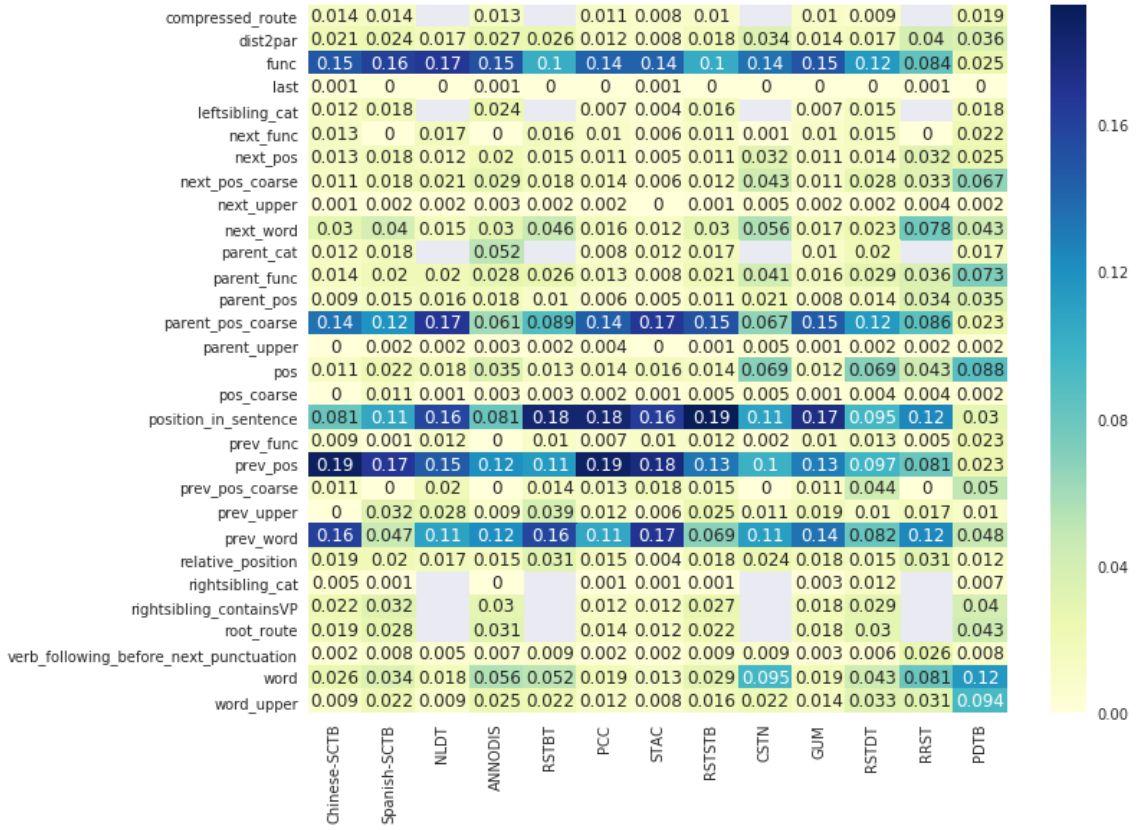

Figure 2: Information gain for the RandomForest classifier.

a relatively important role, while for others this is much less informative. Looking specifically at data sets from the same language (allowing to factor out language differences, and in some cases maintaining a genre difference only), the most notable differences is the informativeness of the part-of-speech tag when comparing *GUM*, *RSTDT* and *STAC* (i.e. it is informative for *RSTDT* but not for the other English data sets). The binary feature for last word in the sentence is partially encoded by relative sentence position as well and in general is very uninformative (with no information gain for most data sets). Surprising is the difference in granularity for the part-of-speech tags. We included both the fine-grained and the coarse tag, suspecting that the fine-grained one may exhibit too much variation for the classifier to pick up on. This does not seem to be the case for the part-of-speech tag of the word itself and that of the previous word. For the parent however, the coarse part-of-speech tag is generally more informative than its more fine-grained version. The data sets in Figure 2 are ordered by size (smallest to largest), but it does not seem to be the case that certain features become more or less informative once data sizes increase.

Note that we largely leave the PDTB, by far the biggest resource of them all, out of this discussion (and consequently also out of Figure 1) due to its different nature of segmentation (at least in the context of the shared task). The task description here notes that for the PDTB-style corpora, "the task is to identify the spans of discourse connectives that explicitly identify the existence of a discourse relation." While this sounds like the task is about discourse connective identification, we note that the data set as published in the shared task includes many instances of words that would not be considered connectives by the usual definitions, such as verbs, nouns and in general includes many alternative lexicalisations. In this case, the baseline scores exceptionally low, as it makes little sense to assume a connective at the start of every sentence. Figure 2 also shows that all the syntactic features[8] add little information for the *PDTB*, and the focus on the surface form could be evidence that the classifier just tries to memorise the words

[8] Although the part-of-speech tag, which can be seen as some kind of syntactic information, does seem to be informative.

as the only thing to go on. However, because we did not investigate this in much detail, we intentionally and equally leave it out of the discussion regarding feature information gain. While due to its size, this data set can potentially contribute a lot to machine-learning based approaches, we argue that a higher degree of unification in the segmentation procedure should be realised before cross-fertilisation can happen. Even though standardising the segmentation task in a theory-neutral way is at the core of the shared task at hand, we found that a better definition and corresponding annotated data set would be needed before reliable classifiers can be constructed. For an idea of connective disambiguation scores on the *PDTB*, we refer to *reference_anonymised*.

We experimented with multilingual word embeddings (Conneau et al., 2017) to have a shared representation for the word and used the syntax features from the dependency layer (as this follows the Universal Dependencies scheme). This allows training on the entire collection (all data sets), and evaluating on just the development set of interest. This however did not improve results compared to using just the data set's corresponding training set.

It seems then that the language usage (i.e. factors like domain, complexity and target audience) plays a more important role in the task of discourse segmentation than the language (i.e. Spanish, Dutch or English for example) in which it is written does. This is also noted by Iruskieta et al. (2016) who look at Basque and Brazilian Portuguese specifically, but equally include and compare texts from different genres. Text from a particular genre from language can thus potentially serve as training data for text from that same genre, but in a language for which no training data for this task is available. We consider further investigation into this direction, adhering to a genre-based distinction rather than a language-based one, the most important pointer to future work and the most promising for performance improvement. First concrete steps in this direction can be the grouping of the data sets included in our experiments in combination with the multilingual word embeddings approach mentioned above.

With regard to the unification of different frameworks, as demonstrated in our experiments, the same systems that work well for EDU segmentation perform very poor for the PDTB-style segmentation defined for the purpose of this shared task. Since shallow annotations are typically easier to obtain and therefore their corresponding corpora can grow larger more easily, the mapping of segments and their properties from a (shallow) theoretical framework (i.e. PDTB) to another (i.e. RST, SDRT or CCR) is a promising direction, but also one that needs more research. Earlier work in this direction ((Demberg et al., 2017), (Sanders et al., 2018) and (Scheffler and Stede, 2016)) may help in the definition of a unifying minimal segment for future attempts at the segmentation task.

7 Conclusion & Outlook

We perform the task of discourse segmentation for various languages, genres and data sets, focusing on segmenting a text into EDUs. Experimenting with 14 data sets from 9 languages representing a variety of domains, we try a RandomForest classifier and an LSTM classifier, and use the same setup for different languages and domains. With the results of the two approaches being close together and no clear winner emerging, the main take-away is that not the language, but the genre seems the most reliable indicator of segmentation performance. We consider more research into genre differences with respect to discourse segmentation the most important suggestion for future work. In addition, while a large corpus with shallow annotations like the PDTB has a lot of potential for improving machine-learning based approaches, we argue that a more refined, unified notion of a minimal segment is needed for cross-theory segmentation to succeed.

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 323949969. We would like to thank the anonymous reviewers for their helpful comments on an earlier version of this manuscript.

References

Stergos Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Cécile Fabre, Lydia-Mai Ho-Dac, Anne Le Draoulec, Philippe Muller, Marie-Paule Péry-Woodley, Laurent Prevot, Josette Rebeyrolle, Ludovic Tanguy, Marianne Vergez-Couret, and Laure Vieu. 2012. An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages –, Istan-

bul, Turkey. European Language Resources Association (ELRA).

N. Asher, A. Lascarides, S. Bird, B. Boguraev, D. Hindle, M. Kay, D. McDonald, and H. Uszkoreit. 2003. *Logics of Conversation*. Studies in Natural Language Processing. Cambridge University Press.

Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos D. Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *LREC*. European Language Resources Association (ELRA).

Farah Benamara and Maite Taboada. 2015. Mapping different rhetorical relation annotations: A proposal. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 147–152. Association for Computational Linguistics.

Chloé Braud, Maximin Coavoux, and Anders Søgaard. 2017. Cross-lingual RST Discourse Parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 292–304. Association for Computational Linguistics.

Harry Bunt and R. Prasad. 2016. ISO DR-Core (ISO 24617-8): Core Concepts for the Annotation of Discourse Relations. In *Proceedings 10th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 45–54.

Shuyuan Cao, Iria da Cunha, and Mikel Iruskieta. 2018. The RST Spanish-Chinese Treebank.

P.C.F. Cardoso, E.G. Maziero, M.L.C. Jorge, E.M.R. Seno, A. Di Felippo, L.H.M. Rino, M.G.V. Nunes, and T.A.S. Pardo. 2011. CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In *Proceedings of the 3rd RST Brazilian Meeting, pp. 88-105. October 26, Cuiab/MT, Brazil*.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002. RST Discourse Treebank, ldc2002t07.

Christian Chiarcos. 2014. Towards interoperable discourse annotation. discourse features in the ontologies of linguistic annotation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

François Chollet et al. 2015. Keras. `https://keras.io`.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Iria da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. 2011. On the Development of the RST Spanish Treebank. In *Proceedings of the 5th Linguistic Annotation Workshop*, LAW V '11, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vera Demberg, Fatemeh Torabi Asr, and Merel Scholman. 2017. How consistent are our discourse annotations? insights from mapping RST-DT and PDTB annotations. *CoRR*, abs/1704.08893.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Hugo Hernault, Helmut Prendinger, David A. duVerle, and Mitsuru Ishizuka. 2010. HILDA: A Discourse Parser Using Support Vector Machine Classification. *Dialogue and Discourse*, 1(3).

M. Iruskieta, M. Aranzabe, A. Diaz de Ilarraza, I. Gonzalez, I. Lersundi, and O. Lopez de Lacalle. 2013. The RST Basque TreeBank: an online search interface to check rhetorical relations. In *4th Workshop RST and Discourse Studies, 40-49, Sociedad Brasileira de Computacao, Fortaleza, CE, Brasil. October 20-24 (http://encontrorst2013.wix.com/encontro-rst-2013)"*.

Mikel Iruskieta, Gorka Labaka, and Juliano D. Antonio. 2016. Detecting the central units in two different genres and languages: a preliminary study of brazilian portuguese and basque texts. *Procesamiento del Lenguaje Natural*, 56:65–72.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13–24. Association for Computational Linguistics.

Shafiq Joty, Giuseppe Carenini, and Raymond T Ng. 2015. CODRA: A Novel Discriminative Framework for Rhetorical Analysis. 41:3.

Hans Kamp. 1981. A Theory of Truth and Semantic Representation. In J. A. G. Groenendijk, T. M. V. Janssen, and M. B. J. Stokhof, editors, *Formal Methods in the Study of Language*, volume 1, pages 277–322. Mathematisch Centrum, Amsterdam.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430, Stroudsburg, PA, USA. Association for Computational Linguistics.

William Mann and Sandra Thompson. 1988. Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. *Text*, 8:243–281.

Eleni Miltsakaki, Aravind K. Joshi, Rashmi Prasad, and Bonnie L. Webber. 2004. Annotating discourse connectives and their arguments. In *FCP@NAACL-HLT*.

Stephan Oepen, Jonathon Read, Tatjana Scheffler, Uladzimir Sidarenka, Manfred Stede, Erik Velldal, and Lilja Øvrelid. 2016. OPT: OsloPotsdamTeesside—Pipelining Rules, Rankers, and Classifier Ensembles for Shallow Discourse Parsing. In *Proceedings of the CONLL 2016 Shared Task*, Berlin.

T. A. S. Pardo and M. G. V. Nunes. 2008. On the Development and Evaluation of a Brazilian Portuguese Discourse Parser. *RITA*, 15:43–64.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pages 13–16. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of LREC*.

Gisela Redeker, Ildik Berzlnovich, Nynke van der Vliet, Gosse Bouma, and Markus Egg. 2012. Multilayer discourse annotation of a dutch text corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Ines Rehbein, Merel Scholman, and Vera Demberg. 2016. Annotating Discourse Relations in Spoken Language: A Comparison of the PDTB and CCR Frameworks. In *LREC*.

Ted J.M. Sanders, Vera Demberg, Jet Hoek, Merel C.J. Scholman, Fatemeh Torabi Asr, Sandrine Zufferey, and Jacqueline Evers-Vermeul. 2018. Unifying dimensions in coherence relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*, 0(0). Exported from https://app.dimensions.ai on 2019/02/06.

Ted J.M. Sanders, Wilbert P.M.S. Spooren, and Leo G.M. Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse Processes*, 15(1):1–35.

Tatjana Scheffler and Manfred Stede. 2016. Mapping pdtb-style connective annotation to RST-style discourse annotation. In *Proceedings of KONVENS*, Bochum, Germany.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 149–156, Stroudsburg, PA, USA. Association for Computational Linguistics.

Manfred Stede and Arne Neumann. 2014. Potsdam Commentary Corpus 2.0: Annotation for discourse research. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Svetlana Toldova, Dina Pisarevskaya, Margarita Ananyeva, Maria Kobozeva, Alexander Nasedkin, Sofia Nikiforova, Irina Pavlova, and Alexey Shelepov. 2017. Rhetorical relations markers in Russian RST Treebank. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 29–33. Association for Computational Linguistics.

Jianxiang Wang and Man Lan. 2015. A refined end-to-end discourse parser. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 17–24. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM corpus: creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Yuping Zhou and Nianwen Xue. 2015. The Chinese Discourse TreeBank: A Chinese Corpus Annotated with Discourse Relations. *Lang. Resour. Eval.*, 49(2):397–431.